Outcome and disease activity in ankylosing spondylitis : an international study

Chapter 8

SUMMARY AND GENERAL DISCUSSION

Summary and general discussion

To create more uniformity in studies concerning aspects of outcome and disease activity in AS, the international working group on ASsessments in Ankylosing Spondylitis (ASAS) defined 'core sets' for the following three settings: disease controlling anti-rheumatic therapy (DC-ART), symptom modifying anti-rheumatic drugs (SM-ARD)/physical therapy and clinical record keeping[1,2]. The domains for all three core sets are physical function, pain, spinal mobility, spinal stiffness, patient global assessment and fatigue. The core sets for clinical record keeping and DC-ART were extended with the domains acute phase reactants, peripheral joints and enthesis and the core set of DC-ART includes also radiographs of spine and hip. As a follow up of the work of the ASAS working group this thesis focuses on different aspects of disease activity and outcome in AS. The first part of the thesis (chapter 2-5) highlights aspects of disease activity and chapters 6 and 7 focus on outcome measures in AS. Most of the results described and discussed in the chapters are derived from an international observational multicenter project: the Outcome in Ankylosing Spondylitis International Study (OASIS). A total of 217 consecutive outpatients with AS who satisfied the modified New York criteria[3] were included in OASIS. This cross sectional cohort of AS patients is followed longitudinally and patients are derived from outpatient clinics of several university hospitals and general hospitals in three European countries: 137 patients from the university hospital Maastricht and the Maasland hospital Sittard (the Netherlands), 55 patients from the hospital Côchin, Paris (France) and 25 patients from the university hospital Ghent (Belgium). These hospitals are secondary and tertiary referral centers. Approximately two third of the OASIS patients are male, a distribution usually seen in AS populations. At baseline of the study the mean age of the patients was 43 years (SD: 13 years) and the mean disease duration since diagnosis was 11 years (SD: 8 years). 27% of the patients had peripheral arthritis diagnosed by their treating rheumatologist. In each country the same trained person (2 rheumatologists and 1 research nurse) assessed all patients every six months according to a pre-specified protocol for a period of two years. All patients were followed by their rheumatologist, independently of the evaluations of the researchers.

Comparison of two functional indexes in AS

Physical function is both related to disease activity and damage in AS. The ASAS working group has also included physical function, assessed by the Dougados Functional Index (D-FI)[4] or the Bath Ankylosing Spondylitis Functional Index (BASFI)[5], in the core sets for all settings[1]. In **chapter 2**, we compared these two widely applied and validated functional indexes specific for AS. The main purpose of this cross-sectional study was to investigate the relation of BASFI and D-FI to aspects of disease activity and damage, which are both related to physical function. If one of the two indexes would perform better, this could be

selected as the preferred measure to assess physical function. The BASFI consists of 10 questions on a visual analogue scale (VAS), all questions deal with activities of daily living. The final score is the average of the scores of the 10 items. The D-FI consists of twenty 5-point Likert response items, assessing the ability to perform distinct daily activities. The total score (ranging from 0-40) is calculated as the sum of the item scores. Because there is no 'gold standard' for disease activity in AS available we used three external criteria for disease activity: both patient and physician assessment of disease activity on a VAS, (0-10) and the Bath Ankylosing Spondylitis Disease Activity Index (BASDAI, range 0-10)[6]. A global and a detailed radiological scoring system specific for AS: the Bath Ankylosng Spondyltis Radiology Index-spine (BASRI-s, range 0-12)[7,8,9] and the modified Stoke Ankylosing Spondylitis Spine Score (SASSS, range 0-72)[10,11] were chosen as the external criterion for damage. In our analyses we used the most contrasting groups of AS patients in which disease activity on the three disease activity measures was defined as high (score ≥ 6.0) or low disease activity (score ≤ 4.0). Furthermore, Receiver Operator Curves (ROC) were plotted for both functional indexes with the three measures of disease activity.

Results of this study showed relatively low disease activity scores and functional scores in our patient population. A total of 7 BASFI questionnaires showed one or more missing answers versus 28 D-FI questionnaires. As expected both functional indexes were highly correlated with each other (Spearman: 0.89). Correlation of BASFI and D-FI with the disease activity measures were all comparable with the highest correlation for the BASDAI (Spearman: 0.59, 0.57 resp.). ROC's of the two functional indexes and all three disease activity measures showed the best curve with the highest sensitivity and specificity values for both functional indexes and the BASDAI (BASFI: 94%, 87% resp. and D-FI: 93%, 79% resp.). The cutoff values to determine high versus low disease activity were considerable higher for the BASFI (_ 40% of the full scale) compared to D-FI (_ 20 % of full scale). However, using these cutoff values showed that considerable percentages of patients were misclassified (12-30%) as having high or low disease activity if solely based on their functional scores. The proportion of misclassified patients was lowest for the BASFI cutoff values in combination with disease activity measured with the BASDAI (12%). So, disease activity assessed with BASDAI comes most closely to the disease activity aspect of function assessed with BASFI. A reason for this may be that both BASDAI and BASFI are developed by the same research group and completely patient reported. Physical function in AS is not solely based on disease activity therefore it could not be expected that all patients were correctly classified. Of the external criteria chosen for damage, BASRI-s appears to give a relatively higher median score for radiological damage than the modified SASSS method: 7.0 (range 2-12) and 12.0 (range 0-72) respectively. Correlation of both functional indexes with BASRI-s and SASSS were about similar, 0,42 and 0.36 respectively.

Although, in this cross-sectional study, BASFI seems to perform slightly better assessing the disease activity aspect and in feasibility (BASFI takes less time to complete and less missing values were found) otherwise no definite choice can be made between BASFI and D-FI based on these results.

The BASFI was developed several years after the D-FI and avoids some presumably redundant items and items assessing symptoms instead of function[12]. Furthermore the BASFI included three items that improve content validity. For two of these items this was proven in the development of the HAQ-s[13]. In case of the third item, concerning physically demanding activities, it was shown in rheumatoid arthritis[14] and osteoarthritis[15] that only this item might discriminate for almost but not perfectly healthy patients who would not score on any other item. At the other hand the BASFI comprises a few items reflecting unusual tasks, and the DFI includes several items covering additional domains not included in the BASFI. A literature review was published including all available studies comparing the performance of BASFI and D-FI. The one study concerning a head to head comparison of the two indexes showed that both instruments were able to discriminate inpatients from outpatients, but only the BASFI could discriminate the effects of a 3 weeks intensive physical therapy period[12]. In four physical therapy trials the D-FI did not discriminate between the treatment arms while the BASFI could discriminate between treatment arms in two other physical therapy trials[12]. A reason for this could be that the distribution of the D-FI scores show a tendency towards normal scores in most studies where it was used. This may not allow further measurement of improvement in patients with only mild disability. To improve the sensitivity of the D-FI the authors proposed the 5 point Likert response scale instead of the 3 point Likert response scale,

The D-FI discriminated well between treatment arms in all but one SMARD trial and in one DC-ART trial between treatment arms[12]. The BASFI did not discriminate between treatment arms in one of the two SMARD trials where it was used[12]. The one SMARD trial concerning a head-to-head comparison of the two indexes showed that both instruments discriminated equally well between placebo and treatment group[12]. There is no DC-ART study available which shows the results of a direct comparison of the two indexes. The only study that evaluates a conventional DC-ART (salazopyrine) the D-FI was used and in three other studies evaluating the effects of inhibition of tumor necrosis factor α only the BASFI was used[12,16,17,18]. In all these studies the applied instrument discriminated between the treatment arms. Based on all these results there is a slight preference for using the DIFI in SMARD trials whereas the BASFI should be preferred in trials concerning physical therapy. Given the efficacy of biological drugs in the treatment of AS direct comparison of both instruments should be preformed in future DC-ART trials and should include calculation of the effect sizes or standardized response mean of the instruments in these settings,

Acute phase reactants in AS

Both laboratory blood test, Erythrocyte Sedimentation Rate (ESR, mm/hr) and C-Reactive Protein (CRP, mg/l), are frequently used to evaluate disease activity in patients with AS. Assessments of acute phase reactants is also a recommended core set endpoint for DC-ART and clinical record keeping by the ASAS Working Group[1,2]. It is well known that the

mean values of these two acute phase reactants are considerably lower for patients with AS in comparison with patients suffering from rheumatoid arthritis[19]. **Chapter 3** describes a study which was conducted to determine whether ESR or CRP is more appropriate in measuring disease activity in AS. Because there might exist differences with respect to ESR and CRP in AS patients depending on the clinical disease presentation, we divided our study population into two groups: patients with only spinal involvement (n=149) and patients with active peripheral arthritis and/or inflammatory bowel disease (n=42). Since there is no 'gold standard' for disease activity in AS we studied the relationship of CRP and ESR with three substitute clinical disease activity variables following the same methodology and statistical approach as described in the previous study on the comparison of BASFI and DFI. A second aim of this cross-sectional study was to determine if elevated CRP or ESR reflect active disease defined by one of these three selected disease activity variables.

The results showed that in the spinal group the majority of patients have normal values for ESR and CRP whereas the majority of patients in the peripheral arthritis/IBD group have slightly elevated levels for both acute phase reactants. Only for ESR this difference between the two groups was statistically significant. Thirty percent of patients in both disease subgroups showed either an elevated ESR with normal CRP or vice versa and in most of these cases values just above normal were seen in case the acute phase reactant was increased. Overall, relatively low disease activity scores were seen in both study groups. Quite striking is the difference in judgement of disease activity between the physician at one hand and both the patient derived measures (BASDAI and patient assessment of disease activity) on the other hand. This is reflected in very different mean values at baseline of physician assessment of disease activity versus BASDAI and patient assessment of disease activity (for the spinal group: 1.5 versus 3.6/3.9 resp. and for the arthritis/IBD group: 2.5 versus 4.3/4.1 resp.). Also in the spinal group only 3% of patients were classified as having high disease activity according to the judgment of the physician. In contrast, the percentage of patients in the spinal group with high disease activity defined by BASDAI and patient was 11% and 21% resp.. Overall, the ROCs showed low cutoff values for ESR and CRP in both groups. Based on these cutoff values sensitivity and specificity were reasonable with the highest sensitivity for physician assessment of disease activity (100%). Unfortunately the corresponding positive predictive values, which are of importance in clinical practice, were uniformly low with large percentages of misclassified patients. This cross-sectional study showed that neither ESR nor CRP are good reflections of active disease as defined in our study. So, no preference can be made to use either of these acute phase reactants in the assessment of disease activity in AS based on these results. There might be a difference between ESR and CRP in relation to the progression of damage in AS, which needs to be further evaluated. Other studies concerning a head-to-head comparison of ESR and CRP are difficult to interpret since different definitions of active disease activity in AS were used. A literature review was published including all available studies comparing the performance of ESR and CRP[20]. In five of these seven studies ESR

and CRP performed equally and two studies indicated that CRP is more closely related to disease activity[20]. There are two studies indicating that elevated ESR and CRP are more likely in patients with peripheral manifestations of AS[20].

There is no SMARD trial available concerning a direct comparison of ESR an CRP. In the available SMARD trials no discriminant capacity of the acute phase reactants was found[20]. One of these SMARD trial also reports the standardized response mean for CRP which was low[21]. Nine DC-ART trials (one methotrexate trial and eight sulfazalazine trials) showed no significant effect of the tested drugs and possibly therefore CRP or ESR did not discriminate between therapy and placebo[20]. The two DC-ART trials allowing a direct comparison of the two acute phase reactants provided opposite results[20]. In two other studies, evaluating the effects of inhibition of tumor necrosis factor α, both ESR and CRP discriminated between treatment arms[16,17]. Unfortunately the exact effect sizes or standardized response mean for ESR and CRP were not given, although these were high for both (>3). Based on all these results no definite decision can be made to use either ESR or CRP in all three clinical setting defined for AS. Since the promising results of biological drugs in the treatment of AS direct comparison of both acute phase reactants is possible and also calculation of the effect size or standardized response mean is needed. Furthermore, the relation of CRP and ESR with the progression of damage in AS should be investigated in future.

Patient self-assessed joint counts in AS

In **chapter 4** a reliability study on patient self-assessed swollen and painful joints is presented. In AS a minor part of the population has peripheral arthritis (± 25%). Traditionally, either a physician or a well-trained healthcare professional is involved in the clinical assessment of arthritis. If joint counts assessed by a physician could be replaced by patient self-assessed joint counts, this would be an advantage for rheumatologists in clinical practice and especially for researchers. The reliability of patient self-reported joint counts in AS has never been studied. In rheumatoid arthritis, reliability of self-assessed joint counts is studied extensively and the results found are both of good reliability[22,23,24,25,26] and poor reliability[27,28,29,30]. In our study, 217 AS patients were asked to mark their painful and swollen joints on a mannequin designed after the method of Stewart[23] presenting 44 and 40 joints respectively. At the same day, without knowledge of the patient assessment, three investigators (one person for each research center) assessed painful and swollen joints on similar mannequins. Our results showed that, on a group level, there is a consistent difference between the number of tender and swollen joints assessed by the patients and by the physicians with only moderate agreement (Intraclass Correlation Coefficient between 0.51and 0.71) on the total number of joint counts and even poor to moderate agreement (kappa between 0.23 and 0.64) on individual joint counts. Patients scored consistently more tender joints and the physicians scored more swollen joints. Possible explanations for these findings are: (1) AS patients can not differentiate between a tender

joint or pain caused by enthesitis since the entheses are located near the joint and (2) AS patients are not trained to detect a swollen joint. The enthesis index of Mander was only significantly correlated with the total number of swollen joints assessed by the physician. On a patient level the results were even worse shown by visualizing our data with the Bland and Altman method[31]. Self assessed joint counts could still be valuable if patients could differentiate between the absence of arthritis and the presence of mono-, oligo- or polyarthritis. However, for these differentations, perfect concordance rates between patients and physicians were also very low (17%, 17% and 22% resp.). The only good concordance rate found was in case of absence of swollen joints (82%). Consequently, AS patient can tell if their joints are not swollen but in case of swollen joints they are unable to judge the extent of swelling even within the rough categories of mono-, oligo-, or polyarthritis. We did not formally assess test-retest reliability in this study but the results obtained at baseline and after one year follow up showed similar results. Based on the results, joint scores derived by physicians cannot be replaced by patient self-assessed joint counts in AS in general. Only information from patients that there are no swollen joints is sufficiently reliable to be useful.

Disease activity in AS

Since there is extended variety in the clinical picture among different AS patients it is very difficult to define disease activity in AS. Patients may experience axial involvement in all degrees of severity, but may also have extra spinal manifestations. This clinical diversity, both in severity and in localization, makes a high demand on instruments that are supposed to measure disease activity in AS. Furthermore AS patients and rheumatologists seem to have very different understandings about active disease[32]. **Chapter 5** describes on which criteria AS patients and rheumatologists base their judgment on disease activity. Our goal was to explore differences between the patient and the physician perspective of AS disease activity. For this study, data of the OASIS patient cohort were used. The patients in this cohort may be considered to appropriately reflect the spectrum of AS patients seen by rheumatologists, since the patients were included irrespective of gender, age, disease duration, disease severity or disease activity,

In this study disease activity from patient perspective as well as from physician perspective was analysed by dichotomising both patient and physician global disease activity score on a VAS (VAS range: 0 not active and 10 extremely active) into 'high disease activity' (VAS $\geq$ 6.0) and 'low disease activity' (VAS $\leq$ 4.0). Various AS instruments selected by the ASAS working group were assessed every six months for two years. Data reduction of these instruments by principal components analysis (PCA) was performed and distinguished four factors capturing correlated instruments, therefore assumed to measure the same underlying construct: spinal mobility, physician assessments, patient assessments and laboratory assessments (Cronbachs alpha between 0.52 and 0.81; explained variance 61%).

Discriminant function analysis with the factor loadings was performed to discriminate between the low- and high disease activity state for both patient and physician perspective. This analysis showed that the factor patient assessments was most important (pooled correlation: 0.84) in discriminating between low and high disease activity state as defined by the patient. The other three factors contributed marginally (pooled correlation: <0.30). In contrast, the three factors: physician assessments, spinal mobility and laboratory assessments contributed most in discriminating between the two defined levels of disease activity of the physician perspective (pooled correlation: 0.62, 0.48, 0.48 respectively). The factor patient assessments did not contribute at all (pooled correlation: 0.05). The discriminant function analysis of baseline data and the analyses of data from other time points revealed similar information. Multiple regression analysis on the discriminant scores was performed to prioritise the instruments with respect to their contribution to each disease activity perspective. In case of the patient perspective disease activity was best captured by the instruments: 'pain spine', 'BASFI', 'pain joints' and 'fatigue'. The physician perspective was best captured by the instruments: 'cervical rotation', 'swollen joint count', 'CRP' and 'intermalleolar distance'.

According to our results AS patients and their physicians indeed have very different views on what disease activity in AS means. AS patients seem to rate disease activity on the basis of complaints while physicians rate disease activity on the basis of instruments assessing inflammation and disease severity. There are a few more conclusions that can be derived from this study. The BASFI, an index primarily designed to assess function in AS also contributes to disease activity from the patient perspective. It also seems that AS patients base part of their estimation of disease activity on what they are able to physically perform. Overall, AS patients appear to properly distinguish disease activity (defined by them as complaints) from disease severity. Disease activity from patient perspective is not captured by acute phase reactants. The physician judgement of disease activity is based on a combination of constructs including measures that combine information on disease activity and severity. Remarkably, CRP was included as a variable while this information was not available to the investigator at the time the judgement of disease activity was made. At the moment fully patient derived instruments such as BASDAI are combined with physicians' assessment of disease activity and/or elevated CRP are used as a 'gold standard' for the assessment of disease activity in AS for including patients in clinical trials and for the start of anti-TNF therapy in clinical practice. Furthermore, there is still lack of evidence that either patient- or physician derived assessments of disease activity are associated with long-term outcome in AS. This important information is needed to be able to select the instruments that best reflect real disease activity leading to the final outcome.

Since the options for drug therapy in AS are increasing it becomes more important to define measures assessing a uniform construct of disease activity and outcome to be used in clinical trials.

Quality of life in AS

In chronic disabling conditions there is a growing interest in the assessment of quality of life (QoL). Especially in studies designed to assess the impact of new pharmaceutical products or to compare different treatment regimes it is becoming relatively common to measure QoL. Disease specific instruments used to evaluate the course of AS focus predominantly on physical impairment and/or physical functioning. Generic health status instruments are available but currently no disease specific instrument exists for assessing quality of life (QoL) in AS patients.

Chapter 6 describes the development of the Ankylosing Spondylitis Quality of Life questionnaire (ASQoL). Our goal was to produce a valid and reliable AS-specific QoL measure that would be relevant and acceptable to respondents. The ASQoL is a quality of life instrument specific to AS and was developed in parallel in the United Kingdom and the Netherlands. All included AS patients fulfilled the modified New York criteria. The methodology used to develop the ASQoL combines the theoretical strengths of the needs-based quality of life model[33] with the statistical and diagnostic power of the Rasch model[34]. The development of the questionnaire enclosed five stages. The first content of the questionnaire was derived from interviews with 30 patients in the UK and 25 patients in the Netherlands (stage 1). Stage 2 concerned the selection of items and response format which formed the first 41-item draft-questionnaire. To assess face and content validity 15 patient field-test interviews were done in both the UK and NL which lead to a 36-item questionnaire (stage 3). Stage 4 concerned a postal survey in the UK (n = 121) to produce a more efficient version of the ASQoL with Rasch analysis the number of questions was reduced to 26 items. Rasch analysis of data from a final postal survey (UK: n = 164; NL: n = 154) was done to assess scaling properties, reliability, internal consistency and construct validity in each country (stage 5). This analysis showed some item misfit, but showed that items formed a hierarchical order and were stable over time. The problematic items were removed resulting in the 18 items ASQoL. Both language versions of the ASQoL showed excellent internal consistency (Conbach's alpha: 0.89-0.91), test-retest reliability (intraclass correlation coefficient: UK: 0.92; NL: 0.91), and validity. The ASQoL may serve a valuable tool in both clinical settings and research for assessing the impact of AS and its treatment on quality of life from the patients perspective. Independently of this study good reliability, validity and responsiveness of the ASQoL was found in another study comparing disease specific patient assessed measures of health outcome in AS[35]. Since the development of the ASQoL this instrument was also used in two trials. One study evaluates the effects of spa therapy and the two other study evaluates the effects of inhibition of tumor necrosis factor α and both studies show that ASQoL discriminated between treatment arms[36,18]. The standardized response mean (SRM) and effect size (ES) were only calculated in case of the spa therapy trial with moderate responsiveness scores (SRM: 0.24 and ES: 0.22) reflecting the moderate treatment effect. In a study on the effect of etanercept there was a high respons.veness of the ASQoL and at least similar to that of the BASDAI[38].

Radiological scoring methods in AS

Radiological damage is considered as an important outcome in AS. The evaluation of radiological change proves to be very difficult in AS. Changes of the sacroiliac joints (SI) are most frequently scored using the 5 grade New York criteria (0-4)[3] or the nearly similar SI score described by the Stoke group[10]. To evaluate the lumbar and cervical spine in AS there are essentially two different scoring methods. The Bath Ankylosing Spondylitis Radiology Index (BASRI) is a global graded scoring method, quick and easy to perform and developed to score the lateral and anterior-posterior view of the lumbar spine (both views combined (0-4), the lateral view of the cervical spine (0-4)[6] and the hips (BASI-hip, 0-4)[39]. The mean score of the New York scoring method of the SI-joints and the several BASRI scores are also combined in two composite scores: BASRI-spine (2-12) and BASRI-total (2-16)[8,9]. The SASSS for the spine is a more detailed scoring method assessing different features such as squaring, sclerosis and erosions at various locations of each vertebra[10,11]. This method is scored on the lateral view of the lumbar spine on both the anterior and posterior sites of the vertebrae (0-72). The 'modified' SASSS is scored on the lateral view of the lumbar spine only at the anterior site of the vertebrae and on the lateral view of the cervical spine also at the anterior site of the vertebrae (0-72)[40].

In **chapter 7**, we compared reliability and changes over one and two years of all these available radiological scoring methods in AS. Two well trained observers scored sets of radiographs from the OASIS cohort at baseline, one and two years follow up. These sets were scored viewing the radiographs simultaneously (paired) without knowledge of the chronology and in random order. The sets of radiographs available for reliability analyses varied from 136 to 200 depending on the number of missing scores for the various scoring methods.

Our results showed good intra- and interobserver reliability for almost all radiological scoring methods. For categorical data, observer agreement was analyzed with linear weighted kappa statistics and in case of continuous data with the Intraclass Correlation Coefficient (ICC). The combined BASRI scoring methods (BASRI-s and BASRI-t) and especially the SASSS showed excellent reliability (ICC 0.85-0.98). Even with a scoring interval of two years the intraobserver reliability remained very good (ICC 0.85-0.96). The reliability of the relatively new scoring method for the hips (BASRI-h) proved to be good (kappa 0.59- 0.60). Of consideration is that kappa indicates to what extent two observers are capable to perceive differences between radiographs. So kappa often turns out to be relatively low in case of a homogeneous group where every single radiograph receives more or less the same score. This could be an explanation for relatively low intra-and interobserver agreement found for the SI scoring methods (kappa 0.36-0.70). Furthermore, measures that relate observed to expected agreement (such as kappa and ICC) are of limited value in this situation because of high levels of expected agreement. This is also confirmed by the relatively low median scores for the SASSS-spine scoring methods (median SASSS-total 17.5-18.0, median modified-SASSS 16.3-16.8, range 0-72). Furthermore,

the low prevalence of radiological damage in SASSS inflates the ICC statistics resulting in a tendency to overestimate the ICC.

Because of these considerations concerning ICC and kappa statistics we also calculated concordance rates. The results showed that the perfect concordance rates between the two observers were overall low (21-76%). Also the visual presentation of a Bland and Altman method[31] adds to the understanding of continuous data (SASSS method) especially because it visualizes the distribution of the data and outliners over the entire range of observed data. These plots showed a maximum difference of 26 points (possible range 0-72) between both observers,

In our study only BASRI-s and BASRI-t were able to detect change based on a binomial cut-off in a small percentage of patients over a two year period (7.5% and 7.4% resp.). This change could not be identified by the other graded and detailed scoring methods. In case of BASRI-s and BASRI-t observers agreed in up to 52% that no change occurred. Unfortunately we may still conclude that relevant change occurred rarely because observers agreed in only 7.5% of cases that real change of at least 1 grade occurred. However, this might be misleading information as we set a change of 1 grade arbitrarily as a cut off. The calculated smallest detectable difference (SDD) for the SASSS is relatively larger[41]. So the cut off used for SASSS seems to be very strict in comparison to the cut off used for the BASRI. This might be a reason why we were unable to detect changes if we applied the SASSS. Furthermore it could be that observer variation or error cannot be distinguished from radiological progression in our study. Moreover, the use of a binomial cut-off induces considerable loss of information and consequently loss of power to detect differences. Another consideration may be that we followed an unselected group of patients, without a particular request for disease activity. In a group of AS patients selected for high disease activity, the situation might be different.

In this study the reliability of AS scoring methods seems to be moderate till good. Unfortunately the scoring methods were unable to detect change over two-year time reliably in a considerable number of patients under the given scoring conditions (paired reading without knowledge of chronology, results based on average score of two observers, cut-off based on SDD, unselected AS population).

Résumé and perspective

Since there was a great need to create more uniformity in different studies focusing on AS the international ASAS working group was formed and this working group defined domains for three AS core sets (DC-ART, SM-ARD/physical therapy, clinical record keeping). In the past 5 years, as a follow up of the work of the ASAS working group, several study groups worldwide, which focus on disease activity and outcome in AS have done a lot of work. Results presented in this thesis are derived from a large cohort of AS patients followed longitudinally (OASIS) and relate to both aspects of outcome and disease activity. Chapter

6 and 7 are focusing on outcome measures in AS. Chapter 6 describes the development of a valid disease specific quality of life instrument (ASQoL). Chapter 7 describes the comparison of available AS radiological scoring methods. These scoring methods prove to be reliable but none of the methods showed considerable change in two year time. In detecting structural change in AS the role of Magnetic Resonance Imaging (MRI) may be become more clear in near future because with MRI it is possible to assess both features of damage and disease activity of the spine and sacroiliac joints in AS[18,42,43]. In case of conventional radiography in AS the possible influence of aspects such as the knowledge of the chronology of the radiographs on sensitivity to change are under investigation,

Chapter 2-5 highlights aspects of disease activity. The results presented in these chapters are all confirming that measuring aspects of disease activity in AS remains very difficult. Although there has been a lot of effort in studying disease activity in AS there is still no uniform measure which reflects AS disease activity in all its aspects. The acute phase reactants such as ESR and CRP are elevated in a minority of AS patients and of most important consideration is that AS patients and their treating physicians seem to have very different understandings about disease activity. Therefore, until now fully patient derived instruments such as BASDAI in combination with physicians assessment of AS disease activity and/or elevated CRP are used as a 'gold standard' to include patients in clinical trials as well as establish anti TNF α therapy in clinical practice. The main reason for this is the persistent lack of a valid tool which combines all aspects of disease activity in AS. Recently more potent biological drugs such as anti TNF α have come available in the treatment of AS and the effects of these drugs are very promising[16,17,18]. In future studies evaluating the effects of these potent drugs can be used to validate and compare ASAS selected instruments used in the follow-up of AS and hopefully these studies will finally lead to the development of a disease activity measure which reflects AS disease activity in all its aspects.

References

1 van der Heijde D, Bellamy N, Calin A, Dougados M, Khan MA, van der Linden Sj. Preliminary core sets for endpoints in ankylosing spondylitis. J Rheumatol 1997; 24: 2225-9.

2 van der Heijde D, Calin A, Dougados M, Khan MA, van der Linden Sj, Bellamy N. Selection of instruments in the core set for DC-ART, SMARD, physical therapy and clinical record keeping in ankylosing spondylitis. Progress report of the ASAS Working Group. J Rheumatol 1999; 26: 952-4.

3 van der Linden S, Valkenburg HA, Cats A. Evaluation. of diagnostic criteria for ankylosing spondylitis: a proposal for modification of the New York criteria. Arthitis Rheum 1984; 27: 361-8.

4 Dougados M, Gueguen A, Nakache JP, Nguyen M, Amor B. Evaluation of a functional index and an articular index in ankylosing spondylitis. J Rheumatol 1988; 15: 02-7.

5 Calin A, Garett S, Whitelock H, et al. A new approach to defining functional ability in ankylosing spondylitis: the development of the Bath Ankylosing Spondylitis Functional Index. J Rheumatol 1994; 21: 2281-5.

6 Garrett S, Jenkinson T, Kennedy LG, Whitelock H, Gaisford P, Calin A. A new approach to defining disease status in ankylosing spondylitis: the Bath Ankylosing Spondylitis Disease Activity Index. J Rheumatol 1994; 21: 2286-91.

7 Kennedy LG, Jenkinson TR, Mallorie PA, Whitelock HC, Garrett SL, Calin A. Ankylosing spondylitis: the correlation between a new metrology score and radiology. Br J Rheum 1995; 34: 767-70.

8 MacKay K, Mack C, Brophy S, Calin A. The Bath Ankylosing Spondylitis Radiology Index (BASRI). A new validated approach to disease assessment. Arthritis Rheum 1998; 41: 2263-70.

9 Calin A, Makay k, Santos H, Brophy S. A new dimension to outcome. Application of the Bath Ankylosing Spondylitis Radiology Index, J Rheumatol 1999; 26: 988-92.

10 Taylor HG, Wardle T, Beswick EJ, Dawes P. The relationship of clinical and laboratory measurements to radiological change in ankylosing spondylitis. Br J Rheum 1991; 30: 330-5.

11 Dawes PT. Stoke Ankylosing Spondylitis Spine Score. J Rheumatol 1999; 26: 993-6.

12 Ruof J, Stucki G. Comparison of the Dougados Functional Index and the Bath Ankylosing Spondylitis Functional Index. a literature review. J Rheumatol 1999; 26: 955-60.

13 Daltroy LH, Larson MG, Liang MH. A modification of the Health Assessment Questionnaire for the Spondyloarthropathies. J Rheumatol 1990;17(7): 946-50.

14 Sangha O, Büchi S, Klaghofer R, Rau R, Stucki G. Development of a new short instrument to assess functional status in rheumatoid arthritis patients [abstract]. Arthritis Rheum 1997; 40 Suppl: S111.

15 Stucki G, Liang MH, Phillips C, Katz JN. The short-form-36 is preferable to the SIP as a generic health status measure in patients undergoing elective total hip arthroplasty. Arthrtis Care Res 1995; 8:174-81.

16 Gorman JD, Sack KE, Davis JC. Treatment of ankylosing spondylitis by inhibition of tumor necrosis factor. N Engl J Med 2002; 18; 1349-56.

17 Braun J, Brandt J, Listing J, Zink A, Alten R, Golder W, Gromnica-Ihle E, Kellner H, Krause A, Schneider M, Sörensen, Zeidler H, Thriene W, Sieper J. Treatment of active ankylosing spondylitis with infliximab: a randomised controlled multicentre trial. Lancet 2002; 359: 1187-93.

18 Marzo-Ortega H, McGonagle D, O'Connor P, Emery P. Efficacy of Etanercept in the treatment of entheseal pathology in resistant spondylarthropathy. Arthritis Rheum 2001; 44: 2112-17.

19 Wolfe F. Comparative usefulness of C-reactive protein and erythrocyte sedimentation rate in patients with rheumatoid arthritis. J Rheumatol 1997; 24: 1477-85.

20 Ruof J, Stucki G.Validity aspects of Erythrocyte Sedimentation Rate an C-Reactive Protein in ankylosing spondylitis - a literature review. J Rheumatol 1999; 26: 966-70.

21 Calin A, Nakache JP, Gueguen A, Zeidler H, Mielant H, Dougados M. Outcome variables in ankylosing spondylitis: evaluation of their relevance and discriminant capacity. J Rheumatol 1999; 26: 975-9.

22 Escalante A. What do self-administered joint counts tell us about patients with rheumatoid arthritis? Arthritis Care Res 1998; 11: 280-90.

23 Stewart MW, Palmer DG, Knight RG. A self-report articular index measure of arthri-tic activity: investigations of reliability, validity and sensitivity. J Rheumatol 1990; 17: 1011-5.

24 Abraham N, Blackmon D, Jackson JR, Bradley LA, Lorish CD, Alarcon GS. Use of self-administered joint counts in the evaluation of rheumatoid arthritis patients. Arthritis Care Res 1993; 6: 78-81.

25 Stucki G, Stucki S, Bruhlmann P, Maus S, Michel BA. Comparison of the validity and reliability of self-reported articular indices. Br J Rheumatol 1995; 34: 760-6.

26 Wong AL, Wong WK, Harker J, Sterz M, Bulpitt K, Park G, et al. Patient self-report tender and swollen joint counts in early rheumatoid arthritis. J Rheumatol 1999; 26: 2551-61.

27 Prevoo ML, Kuper IH, van't Hof MA, van Leeuwen MA, van de Putte LB, van Riel PL. Validity and reproducibility of self-administered joint counts. A prospective longitudinal follow up study in patients with rheumatoid arthritis. J Rheumatol 1996; 23: 841-5.

28 Hanly JG, Mosher D, Sutton E, Weerasinghe S, Theriault D. Self-assessment of disease activity by patients with rheumatoid arthritis. J Rheumatol 1996; 23: 1531-8.

29 Alarcon GS, Tilley BC, Li SH, Fowler SE, Pillemer SR. Self-administered joint counts and standard joint counts in the assessment of rheumatoid arthritis. J Rheumatol 1999; 26: 1065-7.

30 Calvo FA, Calvo A, Berrocal A, Pevez C, Romero F, Vega E, et al. Self-administered joint counts in rheumatoid arthritis: Comparison with standard joint counts. J Rheumatol 1999; 26: 536-9

31 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986; 1: 307-10.

32 Spoorenberg A, van der Heijde D, de Klerk E, Dougados M, de Vlam K, Mielants H, van de Tempel H, van der Linden Sj. Relative value of erythrocyte sedimentation rate and C-reactive protein in assessment of disease activity in ankylosing spondylitis. J Rheumatol 1999; 26: 980-4.

33 Hunt SM, McKenna SP. The QLDS: A scale for the measurement of quality of life in depression. Health Policy 1992; 22: 307-19.

34 Rasch G. Probabilistic Models for some intelligence and attainment tests. Chicago: University of Chicago Press, 1980.

35 Haywood KL, Garratt AM Jordan K, Dziedzic K, Dawes PT. Disease-specific, patient-assessed measures of health outcome in ankylosing spondylitis: reliability, validity and responsiveness. Rheumatol 2002; 41: 1295-1302.

36 van Tubergen A, Landewé R, van der Heijde D, Hidding A, Wolter N, Asscher M, Falkenbach A, Genth E, Goei Thé H, van der Linden Sj. Combined spa-exercise therapy is effective in patients with ankylosing spondylitis: a randomized controlled trial. Arthritis Care Res 2001; 45: 430-8.

37 van Tubergen A, Landewé R, Heuft-Dorenbosch L, Spoorenberg A, van der Heijde D, van der Tempel H, van der Linden Sj. Assessment of disability with the WHODAS II in patients with ankylosing spondylitis. Ann Rheum Dis 2003; 62: 140-5.

38 Marzo-Ortega H, McGonagle D, Emery P: Etanercept treatment in resistant spondyloarthropathy: Imaging, duration of effect and efficacy on reintroduction. Clin Exp Rheumatol 2002; Suppl 28: S175-7.

39 MacKay K, Brophy S, Mack C, Doran M, Calin A. The development and validation of a radiographic grading system for the hip in Ankylosing Spondylitis: the Bath Ankylosing Spondylitis Radiology Hip Index. J Rheumatol 2000; 27: 2866-72.

40 Creemers MCW, Franssen MJAM, van 't Hof MA, Gribnau FWJ, van de Putte LBA, van Riel PLCM. A radiographic scoring system and identification of variables measuring structural damage in Ankylosing Spondylitis [thesis]. 1994; University of Nijmnegen, The Netherlands.

41 Lassere M, Boers M, van der Heijde D, et al. Smallest detectable difference in radiological progression. J Rheumatol 1999; 26: 731-9.

42 Braun J, Sieper J, Bollow M. Imaging of sacroiliitis. Clin Rheumatol 2000; 19: 51-7.

43 Braun J, Baraliakos X, Golder W, Brandt J, Rudwaleit M, Listing J, Bollow M, Sieper J, van der Heijde D. Magnetic resonance imaging examinations of the spine in patients with ankylosing spondylitis, before and after successful therapy with infliximab: evaluation of a new scoring system. Arthritis Rheum 2003; 48: 1126-36

Chapter 9

SAMENVATTING EN DISCUSSIE

Samenvatting en discussie

Een internationale werkgroep (ASsessment in Ankylosing Spondylitis, ASAS) heeft sets met kernpunten gedefinieerd voor wetenschappelijk onderzoek bij de ziekte van Bechterew (ofwel spondylitis ankylopoëtica) met als doel meer uniformiteit te scheppen in wetenschappelijk onderzoek naar ziekte-activiteit en ziekte-uitkomst bij deze aandoening. Deze sets met kernpunten werden gedefinieerd voor drie onderzoeksettings: therapie die het beloop van de ziekte beïnvloedt (disease controlling anti-rheumatic therapy, DC-ART), medicatie en therapie die de symptomen van de ziekte beïnvloeden (symptom modifying anti-rheumatic drugs, SM-ARD/physical therapy) en reguliere behandeling (clinical record keeping). De volgende domeinen werden voor alle drie onderzoeksettings geselecteerd: lichamelijk functioneren, pijn, mobiliteit en stijfheid van de wervelkolom, globale indruk van de patiënt tav ziekte en moeheid. De domeinen voor 'clinical record keeping' en 'DC-ART' zijn uitgebreid met acute fase reacties, perifere gewrichten en peesaanhechtingen (enthesis), 'DC-ART' bevat ook de domeinen: röntgen onderzoek van wervelkolom en heupen.
In navolging van het werk van de ASAS werkgroep heeft dit proefschrift vooral betrekking op verschillende aspecten van ziekte-activiteit en ziekte-uitkomst bij de ziekte van Bechterew. Het eerste deel van het proefschrift (hoofdstuk 2-5) gaat vooral over ziekte-activiteit terwijl de hoofdstuken 6 en 7 zich met name richten op ziekte-uitkomst. Bijna alle resultaten beschreven in dit proefschrift zijn afkomstig van een internationaal observationeel 'multicenter' onderzoek: 'Outcome in Ankylosing Spondylitis International Study' (OASIS). In deze studie werden 217 opeenvolgende poliklinische patiënten met de ziekte van Bechterew geïncludeerd. Dit cross-sectioneel cohort van Bechterew patiënten voldeed aan de gemodificeerde 'New York' criteria en werd longitudinaal gevolgd in verschillende academische- en perifere ziekenhuizen in Europa. Honderdzevenendertig patiënten zijn afkomstig uit het academisch ziekenhuis Maastricht en het Maasland ziekenhuis in Sittard (Nederland), 55 patiënten uit het Hôpital Cochin in Parijs (Frankrijk) en 25 patiënten uit het academisch ziekenhuis Gent (Belgie). Al deze ziekenhuizen zijn secondaire en/of tertiaire referentie centra. Overeenkomstig met andere Bechterew populaties is ongeveer twee derde van de OASIS patiënten van het mannelijk geslacht. Aan het begin van de OASIS studie was de gemiddelde leeftijd van de patiënten 43 jaar (SD 13 jaar) en hadden de patiënten een gemiddelde ziekte duur van 11 jaar (SD 8 jaar). Bij 27% van de patiënten werd een perifere artritis vastgesteld door de behandelend reumatoloog. In elk van de drie participerende landen werden de OASIS patiënten ieder half jaar gedurende 2 jaar onderzocht door steeds dezelfde getrainde persoon (2 reumatologen en 1 onderzoeksverpleegkundige) volgens een vastgesteld protocol. Onafhankelijk van de bevindingen van deze onderzoekers werden de patiënten ook regulier gezien door de behandelend reumatoloog.

De vergelijking van twee indexen voor fysiek functioneren bij de ziekte van Bechterew

Fysiek functioneren is gerelateerd aan ziekte-activiteit en uiteindelijke schade aangericht door de ziekte van Bechterew. De ASAS werkgroep selecteerde fysiek functioneren gemeten met de 'Dougados Functionele Index' (DFI) en de 'Bath Ankyloserende Spondylitis Functionele Index' (BASFI) voor de domeinen van alle drie de onderzoeksettings. In **hoofdstuk 2** werden deze twee veel gebruikte en gevalideerde ziekte specifieke functionele indexen met elkaar vergeleken. Aangezien ziekte-activiteit en schade beide gerelateerd zijn aan fysiek functioneren was het belangrijkste doel van deze cross-sectionele studie om de relatie van BASFI en DFI met deze twee aspecten van de ziekte te onderzoeken. Wanneer de resultaten van deze studie zouden laten zien dat één van deze twee indexen beter zou presteren, dan zou die index uniform gebruikt kunnen worden voor het meten van fysiek functioneren.

De BASFI bestaat uit 10 vragen op een visuele analoge schaal (VAS) en alle vragen betreffen activiteiten uit het dagelijks leven. Het gemiddelde van de scores van de afzonderlijke 10 vragen vormt de uiteindelijke score. De DFI bestaat uit twintig 5 punts Likert respons vragen over het in staat zijn om verschillende dagelijkse activiteiten uit te voeren. De totale score (range van 0-40) wordt berekend door de som van de afzonderlijke vragen te berekenen. Aangezien er geen 'gouden standaard' bestaat voor het meten van ziekte-activiteit bij de ziekte van Bechterew hebben we voor ziekte-activiteit drie meetinstrumenten gekozen: ziekte-activiteit aangegeven op een visuele analoge schaal (VAS, 0-10 cm) door de arts en de patiënt en de Bath Ankylosing Spondylitis Disease Activity Index (BASDAI, range 0-10). Als meetinstrumenten voor schade werden een globale en een gedetailleerde röntgen scoringsmethode specifiek voor het scoren van de wervelkolom bij de ziekte van Bechterew gekozen: de Bath Ankylosing Spondylitis Radiology Index-spine (BASRI-s, range 0-12) en de gemodificeerde Stoke Ankylosing Spondylitis Spine Score (SASSS, range 0-72). Voor de analyses hebben we de twee meest contrasterende groepen patiënten gebruikt: Bechterew patiënten waarbij ziekte activiteit hoog (score $\geq$ 6.0) en laag (score $\leq$ 4.0) was volgens de drie ziekte-activiteit instrumenten. Verder werden er Receiver Operator Curves (ROC) gemaakt voor beide functionele indexen versus de drie ziekte-activiteit instrumenten.

In onze patiënten populatie werden relatief lage waarden voor de functionele indexen en de ziekte-activiteit instrumenten gevonden. Er waren 7 BASFI vragenlijsten versus 28 DFI vragenlijsten met één of meer niet ingevulde vragen. Zoals verwacht waren beide functionele indexen hoog met elkaar gecorreleerd (Spearman 0.89). De correlaties van de BASFI en DFI met de ziekte-activiteit instrumenten was vergelijkbaar voor beide indexen waarbij correlatie met de BASDAI het hoogst was (respectievelijk 0.59 en 0.57). ROC curven voor elk van de twee functionele indexen met de drie ziekte- activiteit instrumenten liet de beste curve met de hoogste sensitiviteit en specificiteit zien voor beide functionele indexen versus de BASDAI (BASFI: respectievelijk 94% en 87% en DFI: respectievelijk 93%

en 79%) De afkappunten voor het onderscheiden van lage en hoge ziekte-activiteit waren duidelijk hoger voor de BASFI (± 40% van de volledige schaal) in vergelijking met de DFI (± 20 % van de volledige schaal). Wanneer we echter deze afkappunten gebruiken om onderscheid te maken tussen hoge en lage ziekte-activiteit gebaseerd op de functionele scores worden er behoorlijk hoge percentages patiënten fout geclassificeerd (12-30%). Het percentage fout geclassificeerde patiënten was het laagste voor de BASFI afkappunten in combinatie met ziekte-activiteit gemeten met de BASDAI (12%).

Ziekte-activiteit gemeten met de BASDAI komt dus het dichtst bij het ziekte-activiteit aspect van de BASFI. Een reden hiervoor zou kunnen zijn dat beide indexen geheel door de patiënt gerapporteerd worden en door dezelfde onderzoeksgroep ontwikkeld zijn. Aangezien fysiek functioneren bij de ziekte van Bechterew niet alleen bepaald wordt door ziekte-activiteit kan ook niet verwacht worden dat alle patiënten goed geclassificeerd werden. Ten aanzien van de instrumenten gekozen voor het bepalen van schade laat de BASRI-s een relatief hogere mediaan score voor radiologische schade zien dan de gemodificeerde SASSS: respectievelijk 7.0 (range 2-12) en 12.0 (range 0-72). Correlaties van beide functionele indexen met BASRI-s en gemodificeerde SASSS waren vergelijkbaar (respectievelijk 0.42 en 0.36).

In deze cross-sectionele studie lijkt de BASFI iets beter te presteren dan de DFI met betrekking tot aspecten van ziekte-activiteit en uitvoerbaarheid (BASFI kost minder tijd om in te vullen en er waren minder ontbrekende antwoorden). Op basis van de resultaten van dit onderzoek is er anderszins geen voorkeur ten aanzien van BASFI en DFI uit te spreken. De BASFI werd enkele jaren na de DFI ontwikkeld en vermijdt een aantal vermoedelijk overtollige vragen en vragen met betrekking op symptomen in plaats van fysiek functioneren. Verder bevat de BASFI drie vragen die de validiteit van de inhoud verbeteren. Voor twee van deze vragen is dit bewezen bij de ontwikkeling van de HAQ-s. Bij patiënten met reumatoïde artritis en artrose werd aangetoond dat alleen de derde van deze vragen, betreffende fysiek belastende activiteiten, kon discrimineren tussen bijna maar niet geheel gezonde patiënten die niet scoorden op één van de andere vragen.

Aan de andere kant bevat de BASFI vragen over ongewone taken en de DFI bevat diverse vragen met betrekking op verschillende aanvullende domeinen die niet in de BASFI geïncludeerd zijn. Er is een literatuur overzicht gepubliceerd van alle bestaande studies met betrekking tot de vergelijking van de BASFI en DFI. De enige studie met een rechtstreekse vergelijking van de twee functionele indexen liet zien dat beide indexen onderscheid konden maken tussen poliklinische en klinische patiënten maar alleen de BASFI kon de effecten van een intensieve fysiotherapie gedurende drie weken aantonen. In zes studies met betrekking tot fysiotherapie was de DFI vier maal niet in staat onderscheid te maken tussen de verschillende behandelingen terwijl in de twee andere studies de BASFI wel kon discrimineren tussen de verschillende behandelingen. Een reden hiervoor kan zijn dat in de studies waar de DFI gebruikt werd de scores van de DFI vragen dicht bij de normale waarden lagen. Het gevolg hiervan kan zijn dat het niet mogelijk is verder te verbeteren wanneer er sprake is van alleen milde invaliditeit. Om de sensitiviteit van de DFI te verbeteren stelden de auteurs voor de vragen om te zetten van een drie punt Likert

respons schaal naar een vijf punt Likert respons schaal.
De DFI discrimineerde goed tussen de verschillende behandelingen in op één na alle SMARD trials en in één DC-ART studie waarbij deze index werd gebruikt. De BASFI discrimineerde tussen de verschillende behandelingen in één van de twee SMARD trials waarbij de index werd gebruikt. De enige SMARD trial waarbij een rechtstreekse vergelijking van de twee indexen is gedaan liet zien dat beide indexen even goed discrimineerden tussen placebo en behandelgroep. Er is geen DC-ART studie beschikbaar die de BASFI en de DFI rechtstreeks vergelijkt. In de enige studie die een conventionele DC-ART (sulfasalazine) evalueerde werd de DFI gebruikt en in drie andere studies waar de effecten van anti tumor necrosis factor (TNF) a werden geëvalueerd werd de BASFI gebruikt. In al deze studies discrimineerde de gebruikte index goed tussen de verschillende behandelarmen.
Samenvattend lijkt er een lichte voorkeur te bestaan om de DFI te gebruiken in SMARD trials en de BASFI in trials met betrekking tot fysiotherapie. Gezien de goede effecten van biologische geneesmiddelen zoals anti-TNF a bij de behandeling van de ziekte van Bechterew is een rechtstreekse vergelijking van beide functionele indexen in toekomstige DC-ART trials goed mogelijk; daarbij zullen dan ook 'effect sizes' en/of 'standardised reponse mean' van beide indexen berekend moeten worden.

Bezinking (BSE) versus C-reactive protein (CRP) bij de ziekte van Bechterew

De acute fase bloed testen, BSE (mm/uur) en CRP (mg/l), worden beide regelmatig gebruikt ter evaluatie van de ziekte activiteit bij Bechterew patiënten. Het bepalen van acute fase reacties is door de ASAS werkgroep voorgesteld als een domein voor de onderzoeksettings DC-ART trial en reguliere behandeling. Het is bekend dat de gemiddelde waarde van deze twee acute fase reacties laag zijn bij patiënten met de ziekte van Bechterew in vergelijking met patiënten met reumatoïde artritis.
Hoofdstuk 3 van dit proefschrift beschrijft een onderzoek met als doel te bepalen of er onderscheidt gemaakt kan worden tussen BSE en CRP voor het meten van ziekte-activiteit bij de ziekte van Bechterew. Gezien er verschillen kunnen bestaan in het klinisch beeld bij Bechterew patiënten ten aanzien van BSE en CRP is onze patiënten populatie in twee groepen verdeeld: patiënten met alleen spinale betrokkenheid (n=149) en patiënten met ook actieve perifere artritis en/of inflammatoire darmziekte (n=42). Zoals eerder werd aangegeven bestaat er geen 'gouden standaard' voor het meten van ziekte-activiteit bij de ziekte van Bechterew. We bestudeerde daarom de relatie van BSE en CRP met drie vervangende klinische ziekte-activiteit instrumenten volgens dezelfde methodologie en statische procedure als beschreven in de vorige studie met betrekking tot de vergelijking van BASFI en DFI. Een tweede doel van deze studie was na te gaan of hoge waarden voor BSE en CRP de mate van ziekte-activiteit reflecteren gedefinieerd door de drie geselecteerde ziekte-activiteit instrumenten.

De resultaten lieten zien dat in de groep met alleen spinale betrokkenheid de meeste patiënten normale waarden voor BSE en CRP hadden. Daarentegen lieten de Bechterew patiënten met ook perifere artritis en/of inflammatoire darmziekte licht verhoogde waarden voor BSE en CRP zien. Allen voor de BSE was dit verschil tussen de twee subgroepen statistisch significant. Dertig procent van de patiënten in beide subgroepen hadden een verhoogde BSE en een normale CRP of vice versa en in de meeste van deze gevallen werden waarden van net boven de normale grens gevonden als de acute fase reactie verhoogd was. Over het algemeen werden er in beide subgroepen relatief lage scores voor de ziekte-activiteit instrumenten gevonden. Opvallend is het verschil in beoordeling van ziekte-activiteit door de arts (VAS) aan de ene kant en de twee op de patiënt gebaseerde instrumenten aan de andere kant (BASDAI en ziekte-activiteit aangegeven op een VAS door de patiënt). Aan het begin van de studie werd dit weergegeven in zeer verschillende gemiddelde waarden van ziekte-activiteit aangegeven door de arts versus BASDAI en ziekte-activiteit aangegeven door de patiënt (voor de groep met spinale betrokkenheid: respectievelijk 1.5 versus 3.6 en 3.9 en voor de groep met perifere artritis en/of inflammatoire darmziekte: respectievelijk 2.5 versus 4.3 en 4.1). In de groep met alleen spinale betrokkenheid had maar 3% van de patiënten hoge ziekte activiteit beoordeeld door de arts in tegenstelling tot ziekte activiteit beoordeeld door de patiënt en BASDAI (respectievelijk 21% en 11%). Over het algemeen lieten de ROC curven lage afkappunten zien in beide subgroepen voor zowel BSE als CRP. Gebaseerd op deze afkappunten waren sensitiviteit en specificiteit redelijk met de hoogste sensitiviteit (100%) voor ziekte- activiteit aangegeven op een VAS door de arts. Helaas waren de bijbehorende positief voorspellende waarden, die belangrijk zijn in de klinische praktijk, laag met hoge percentages fout geclassificeerde patiënten. Deze cross-sectionele studie laat zien dat zowel BSE als CRP niet overeenkomen met ziekte-activiteit zoals gedefinieerd in deze studie. Op basis van deze resultaten kan er dus geen duidelijke voorkeur worden gegeven aan één van deze twee acute fase reacties. In relatie met progressie van schade bij de ziekte van Bechterew zou er wel een duidelijk verschil kunnen bestaan tussen BSE en CRP en dit zal in de toekomst ook verder geëvalueerd moeten worden.

Overige studies die een rechtstreekse vergelijking van BSE en CRP laten zien zijn moeilijk te interpreteren doordat er steeds verschillende definities voor ziekte-activiteit worden gebruikt. Er is een literatuuroverzicht gepubliceerd van alle bestaande studies met betrekking tot de vergelijking van BSE en CRP bij de ziekte van Bechterew. In vijf van deze zeven studies werd er geen verschil gevonden tussen BSE en CRP en de resultaten van de twee andere studies geven aan dat CRP beter gerelateerd is aan ziekte activiteit dan BSE. Twee studies geven aan dat verhoogde BSE en CRP vaker gezien worden in Bechterew patiënten met perifere manifestaties van de ziekte. BSE en CRP werden niet rechtstreeks vergeleken in een SMARD trial. In de beschikbare SMARD trials werd geen onderscheidend vermogen gevonden van BSE en CRP tussen de verschillende behandelingen. Eén van deze SMARD trials rapporteerde een lage 'standardised response mean' voor CRP. Negen DC-ART trials (één methotrexaat trial en acht sulfasalazine trials) lieten geen significant effect

van de gebruikte medicatie zien en waarschijnlijk is hierdoor ook geen discriminerend vermogen van BSE en CRP zichtbaar tussen de geteste medicatie en placebo. De twee DC-ART trials met een rechtstreekse vergelijking van de twee acute fase reacties lieten tegengestelde resultaten zien. In twee studies naar de effecten van anti-TNF a konden zowel BSE als CRP erg goed onderscheid maken tussen de verschillende behandelingen. Alhoewel de 'effect sizes' en 'standardised response mean'voor BSE en CRP hoog waren in deze studies (>3) werden de exacte waarden hiervan helaas niet gerapporteerd. Op basis van deze resultaten kan voor geen van de drie onderzoeksettings een definitieve keus gemaakt worden tussen BSE en CRP. Gezien de veelbelovende resultaten van biologische geneesmiddelen zoals anti-TNF alfa bij de behandeling van de ziekte van Bechterew is er een rechtstreekse vergelijking van BSE en CRP mogelijk waarbij door berekening van 'effect sizes' en/of 'standardised reponse mean' van beide acute fase reacties een betere vergelijking mogelijk is. Tevens zal in de toekomst de relatie van BSE en CRP met de progressie van schade bij de ziekte van Bechterew onderzocht moeten worden.

Patiënt gerapporteerde gewricht scores bij de ziekte van Bechterew

In **hoofdstuk 4** wordt een studie gepresenteerd naar de betrouwbaarheid van het rapporteren van pijnlijke en gezwollen gewrichten door de patiënt zelf. Een klein gedeelte van de Bechterew patiënten heeft een perifere artritis (± 25%). Normaal gesproken wordt artritis klinisch gediagnostiseerd door een arts of een goed getrainde verpleegkundige. Wanneer de gewrichtsscores aangetoond bij lichamelijk onderzoek door de arts vervangen kunnen worden door gewrichtsscores door de patiënt zelf zou dit een groot voordeel betekenen voor reumatologen en in het bijzonder klinisch onderzoekers. De betrouwbaarheid van door de patiënten zelf gerapporteerde gewrichtsscores bij de ziekte van Bechterew is nooit eerder onderzocht. Bij patiënten met reumatoïde artritis is de betrouwbaarheid van patiënt gerapporteerde gewrichtsscores uitgebreid onderzocht en de resultaten lieten zowel hoge als lage betrouwbaarheid zien. In onze studie werd aan 217 Bechterew patiënten gevraagd hun pijnlijke en gezwollen gewrichten aan te kruisen op een mannequin, ontwikkeld volgen de methode van Stewart waarop respectievelijk 44 en 40 gewrichten aangekruist kunnen worden. Op dezelfde dag maar zonder dat de resultaten van de patiënten bekend waren rapporteerden drie onderzoekers (één persoon voor elk onderzoekscentrum) de pijnlijk en gezwollen gewrichten van deze patiënten op vergelijkbare mannequins. De resultaten lieten zien dat er op groepsniveau een consistent verschil was tussen het aantal gerapporteerde pijnlijke en gezwollen gewrichten door patiënten en onderzoekers met een bijbehorende matige overeenstemming (Intraclass Correlatie Coëfficiënt tussen 0.51 en 0.71) van het totaal aantal pijnlijke en gezwollen gewrichten en zelfs een slechte tot matige overeenstemming (kappa tussen 0.23 en 0.64) van de individuele gewrichtsscores. Patiënten scoorden consistent meer pijnlijke gewrichten en de onderzoekers scoorden meer gezwollen gewrichten. Mogelijke verklaringen hiervoor

zijn : (1) Bechterew patiënten kunnen niet goed differentiëren tussen een pijnlijk gewricht en pijn veroorzaakt door enthesitis gezien de aanhechtingen vlak bij het gewricht gelokaliseerd zijn en (2) Bechterew patiënten zijn niet getraind om gezwollen gewrichten te detecteren. De enthesis index volgens Mander was alleen significant gecorreleerd met het totaal aantal gezwollen gewrichten gerapporteerd door de onderzoekers. Op patiëntniveau waren de resultaten zelfs nog slechter en dit is duidelijk zichtbaar gemaakt in Bland en Altman plots. Door patiënt gerapporteerde gewrichtsscores kunnen nog steeds van waarde zijn als patiënten kunnen differentiëren tussen de aanwezigheid van mono- oligo- en polyartritis. Ook voor deze onderverdeling was volledige overeenstemming tussen onderzoekers en patiënten erg laag (respectievelijk 17%, 17% en 22%). Alleen wanneer er geen sprake was van gezwollen gewrichten was de overeenstemming tussen onderzoekers en patiënten goed (82%). Bechterew patiënten kunnen dus wel oordelen over de afwezigheid van gezwollen gewrichten maar ze zijn niet in staat om om de aanwezigheid van zwelling aan te geven zelfs niet in de grove categorieën van mono- oligo- en polyartritis. In deze studie hebben we officieel geen test-retest betrouwbaarheid onderzocht maar de resultaten verkregen met de baseline en 1 jaars data waren vergelijkbaar. Op basis van deze resultaten kunnen gewrichtsscores gerapporteerd door onderzoekers/artsen dus niet vervangen worden door gewrichtsscores gerapporteerd door de Bechterew patiënten zelf. Alleen informatie van de patiënten ten aanzien van het afwezig zijn van gezwollen gewrichten is voldoende betrouwbaar om bruikbaar te kunnen zijn.

Ziekte-activiteit bij de ziekte van Bechterew

Aangezien er een grote variatie bestaat in het klinische beeld tussen verschillende Bechterew patiënten is het erg moeilijk om ziekte-activiteit te definiëren bij deze ziekte. Patiënten hebben axiale betrokkenheid in verschillende gradaties maar kunnen daarbij ook verschillende extra-spinale manifestaties van de ziekte hebben. Deze klinische diversiteit in zowel ernst als in lokalisatie zorgt ervoor dat instrumenten die gebruikt worden om ziekte-activiteit te meten aan hoge eisen moeten voldoen. Verder hebben Bechterew patiënten en hun behandelend reumatologen zeer uiteenlopende inzichten ten aanzien van ziekte-activiteit. **Hoofdstuk 5** beschrijft op basis van welke criteria Bechterew patiënten en reumatologen ziekte-activiteit beoordelen. Ons doel was om verschillen in inzicht ten aanzien van ziekte-activiteit tussen patiënten en reumatologen te verkennen. Voor deze studie werden data van het OASIS cohort gebruikt. Aangenomen mag worden dat de patiënten uit dit cohort het hele klinische spectrum van Bechterew patiënten omvat dat normaal gesproken door reumatologen gezien worden. De patiënten werden geïncludeerd onafhankelijk van geslacht, leeftijd, duur van de ziekte, ernst van de ziekte en mate van ziekte-activiteit.

In deze studie werd ziekte-activiteit bestudeerd vanuit het perspectief van zowel de patiënt als van de arts. Ziekte-activiteit door de arts en de patiënt aangegeven op een visuele

analoge schaal (VAS range: 0 = niet actief en 10 = zeer actief) werd onderverdeeld in 'hoge' ziekte-activiteit (VAS ≥.6.0) en 'lage' ziekte-activiteit (VAS ≤ 4.0). Meetinstrumenten die door de ASAS werkgroep geselecteerd zijn voor gebruik bij de evaluatie van de ziekte van Bechterew werden iedere zes maanden toegepast over een periode van twee jaar. Datareductie van deze instrumenten werd verricht met behulp van factoranalyse. Dit resulteerde in vier factoren met onderling correlerende meetinstrumenten: 'metingen van de mobiliteit van de wervelkolom', 'metingen door de arts', 'metingen door de patiënt' en 'laboratoriumbepalingen' (Cronbachs alpha tussen 0.52 en 0.81; verklaarde variantie 61%). Er werd aangenomen dat de instrumenten binnen een factor hetzelfde onderliggende construct bepalen. Een discriminantanalyse met de factorwaarden werd verricht om onderscheid te kunnen maken tussen 'lage' en 'hoge' ziekte-activiteit voor zowel het perspectief van de patiënt als dat van de arts. Deze analyse liet zien dat de factor, 'metingen door de patiënt', het meest bijdragend was (gezamenlijke (pooled) correlatie 0.84) in het onderscheid maken tussen de twee niveaus van ziekte-activiteit gedefinieerd door de patiënt. De bijdrage van de andere drie factoren was maar minimaal (pooled correlatie < 0.30). Daarentegen droegen de drie factoren: 'metingen door de arts', 'metingen van de mobiliteit van de wervelkolom' en laboratoriumbepalingen' het meest bij in het discrimineren tussen de twee niveaus van ziekte-activiteit gedefinieerd door de arts (pooled correlatie: respectievelijk 0.62, 0.48 en 0.48). De factor, 'metingen door de patiënt', droeg in het geheel *niet bij (pooled correlatie 0.05). De discriminantanalyse verricht* met baseline data en de data van de andere momenten liet geen verschillen zien. De discriminantscores werden gebruikt voor multipele regressie analyse om de prioriteit van de instrumenten te bepalen ten aanzien van de bijdrage aan ieder perspectief van ziekte-activiteit. Het ziekte-activiteit perspectief van de patiënt werd het beste verklaard door de instrumenten: 'pijn van de wervelkolom', 'BASFI', 'gewrichtspijn', en 'moeheid'. Het perspectief van de arts kwam het beste tot uiting met de instrumenten: 'cervicale rotatie', 'aantal gezwollen gewrichten', CRP' en 'intermalleolaire afstand'. Onze resultaten bevestigen dat patiënten en artsen een heel ander zicht hebben op wat ziekte-activiteit bij de ziekte van Bechterew betekent. Bechterew patiënten beoordelen ziekte-activiteit op basis van klachten terwijl artsen ziekte-activiteit beoordelen op basis van instrumenten die ontsteking en ernst van de ziekte meten.

Uit deze studie volgen nog een aantal andere conclusies. De BASFI, een index primair ontworpen om fysiek functioneren bij Bechterew patiënten te meten, sluit ook aan bij het perspectief van ziekte-activiteit van de patiënt. Het lijkt er tevens op dat Bechterew patiënten hun inschatting van ziekte-activiteit gedeeltelijk maken op basis van wat ze lichamelijk kunnen doen. Over het algemeen kunnen Bechterew patiënten goed onderscheid maken tussen ziekte-activiteit (bepaald door klachten) en de ernst van de ziekte. Ziekte-activiteit vanuit het patiënten perspectief wordt niet bepaald door acute fase reacties. Het oordeel over ziekte-activiteit van de arts wordt gebaseerd op een combinatie van constructen met instrumenten die informatie over ziekte-activiteit en ernst van de ziekte combineren. Opvallend is dat CRP als variabele werd geïncludeerd aangezien de waarde hiervan niet bij de arts bekend was op het moment dat er een oordeel werd

gegeven over ziekte-activiteit. Momenteel worden volledig op het oordeel van de patiënt gebaseerde ziekte-activiteit instrumenten, zoals BASDAI, in combinatie met ziekte activiteit aangegeven door de arts en/of verhoogde CRP gebruikt als 'gouden standaard' voor het meten van ziekte-aciviteit bij de ziekte van Bechterew voor het includeren van patiënten in klinische trials en voor de besluitvorming rondom het starten van anti TNF α therapie in de klinische praktijk. Er is echter nog steeds geen wetenschappelijk bewijs dat ziekte-activiteit instrumenten gebaseerd op het oordeel van de patiënt of de arts geassocieerd zijn met ziekte-uitkomst op de lange duur. Deze belangrijke informatie is nodig om instrumenten te kunnen selecteren die zowel de ziekte-activiteit als de uitkomst van de ziekte het beste weergeven. Aangezien de keuzes voor medicamenteuze behandeling van de ziekte van Bechterew groter worden is het ook belangrijker om instrumenten te definiëren die een eenduidig construct van ziekte-activiteit en ziekte-uitkomst meten zodat deze gebruikt kunnen worden in klinische trials.

Kwaliteit van leven bij de ziekte van Bechterew

Er is een groeiende belangstelling voor het meten van kwaliteit van leven (QoL) bij chronische invaliderende aandoeningen. Met name in studies ontwikkeld om de impact van nieuwe farmaceutische producten te meten of om verschillende behandelmethoden met elkaar te vergelijken wordt steeds vaker kwaliteit van leven gemeten. Ziektespecifieke instrumenten gebruikt om het beloop van de ziekte van Bechterew te evalueren richten zich vooral op fysieke beperkingen en/of fysiek functioneren. Er zijn wel generieke instrumenten die de algehele gezondheid meten maar er bestaat nog geen ziektespecifiek instrument voor het meten van kwaliteit van leven bij de ziekte van Bechterew.

Hoofdstuk 6 beschrijft de ontwikkeling van de Ankylosing Spondylitis Quality of Life vragenlijst (ASQoL) . Ons doel was om een valide en betrouwbaar Bechterew specifiek kwaliteit van leven instrument te ontwikkelen dat relevant en acceptabel is voor respondenten. De ASQoL is gelijktijdig ontwikkeld in Engeland en Nederland. Alle geïncludeerde Bechterew patiënten voldeden aan de gemodificeerde New York criteria. De methodologie gebruikt om de ASQoL te ontwikkelen combineert de theoretische principes van het 'needs-based' kwaliteit van leven model met de statistische en diagnostische kracht van het Rasch-model.

De ontwikkeling van de vragenlijst bestaat uit vijf stadia. De inhoud van de eerste versie van de vragenlijst was afkomstig van interviews met 30 patiënten in Engeland en 25 patiënten in Nederland (stadium 1). Stadium 2 betreft het selecteren van items en antwoord type waarbij de eerste proef vragenlijst met 41 items werd gevormd. Er werden test interviews gedaan met deze proef vragenlijst bij 15 patiënten in Engeland en 15 patiënten in Nederland om informatie over validiteit van vorm en inhoud te verkrijgen. Hierbij ontstond een vragenlijst met 36 items (stadium 3). Stadium 4 was een onderzoek per post in Engeland (n=121) met als doel een efficiëntere versie van de ASQoL te ontwikkelen. Met behulp van Raschanalyse werd het aantal vragen verminderd tot 26 items. Raschanalyse van de data van

een laatste onderzoek per post werd in beide landen (Engeland: n=164; Nederland: n=154) gedaan om eigenschappen van de schaal, betrouwbaarheid, interne consistentie en construct validiteit te meten (stadium 5). Deze analyse liet enkele misplaatste items zien maar liet tevens zien dat de items een hiërarchische volgorde hadden en stabiel waren over de tijd. De problematische items werden verwijderd hetgeen resulteerde in de definitieve ASQoL met 18 items. Beide taal versies van de ASQoL hadden een uitstekende interne consistentie (Cronbach's alfa: 0.89-0.91), test-retest betrouwbaarheid (Intraclass Correlatie Coëfficiënt: Engeland:0.92; Nederland: 0.91) en validiteit. De ASQol kan een waardevol instrument zijn in zowel de klinische situatie als bij wetenschappelijk onderzoek wanneer de invloed van de ziekte van Bechterew en bijbehorende behandeling op kwaliteit van leven gemeten wordt. Onafhankelijk van deze studie werd een goede betrouwbaarheid, validiteit en respons van de ASQoL gevonden in een studie waarbij door de patiënt bepaalde ziekte specifieke instrumenten werden vergeleken. De ASQoL werd sinds de ontwikkeling gebruikt in twee trials. Eén studie evalueert het effect van kuurtherapie en de andere het effect van anti-TNF a therapie. Beide studies laten zien dat de ASQol in staat is te discrimineren tussen de verschillende behandelingen. De 'standardised reponse mean' (SRM) en 'effect size' (ES) werden alleen berekend voor de kuurtherapie trial met matige respons scores (SRM: 0.24 en EF: 0.22) passend bij het matige behandeleffect. Er was een onderling vergelijkbaar hoge respons van de ASQoL en de BASDAI in de studie naar het effect van etanercept.

Radiologische scoringsmethoden bij de ziekte van Bechterew

Radiologische schade is een belangrijke uitkomst maat bij de ziekte van Bechterew. Het is erg moeilijk om deze radiologische schade te evalueren. Veranderingen van de sacroiliacale gewrichten (SI) worden meestal gescoord door gebruik te maken van de New York criteria of de bijna identieke scoringsmethode ontwikkeld door de Stoke groep. Beide SI scoringsmethoden differentiëren 5 graden (range 0-4 voor ieder SI gewricht). Er zijn twee scoringsmethoden om de lumbale en cervicale wervelkolom te scoren. De 'Bath Ankylosing Spondylitis Radiology Index' (BASRI) is een globale, snelle en makkelijke scoringsmethode ontwikkeld voor het scoren van de voor-achterwaartse en de laterale opname van de lumbale wervelkolom (range beide opnames gecombineerd 0-4), de laterale opname van de cervicale wervelkolom (range 0-4) en de heupen (BASRI-h, range 0-4 voor iedere heup; daarna gemiddelde rechter en linker heup). De gemiddelde score van de New York scoringsmethode voor de SI gewrichten gecombineerd met de verschillende BASRI scoringsmethoden vormen twee samengestelde scoringsmethoden: BASRI-'spine' (range 2-12) en BASRI-'total' (range 2-16). De SASSS methode voor de wervelkolom is een meer gedetailleerde scoringsmethode die verschillende aspecten zoals 'squaring', sclerose en erosies op verschillende plaatsen van iedere wervel scoort. Deze methode wordt gescoord op de laterale opname van de lumbale wervelkolom aan de voor- en achterzijde van iedere wervel (SASSS, range 0-72). De 'gemodificeerde' SASSS wordt gescoord op de laterale

opname van zowel de lumbale als cervicale wervelkolom waarbij alleen de voorzijde van iedere wervel gescoord wordt (range 0-72). In **hoofdstuk 7**, werd de betrouwbaarheid en radiologische verandering na 1 en 2 jaar follow-up van al deze beschikbare radiologische scoringsmethoden vergeleken. Twee goed getrainde 'observers' scoorden sets met röntgenfoto's van het OASIS cohort gemaakt op baseline, 1 en 2 jaar follow-up. De röntgenfoto's van één set werden gelijktijdig (gepaard) gescoord zonder kennis van de chronologische volgorde. De volgorde waarin alle sets werd gescoord was willekeurig. Afhankelijk van het aantal missende scores voor iedere methode varieerde het aantal beschikbare sets voor betrouwbaarheidsanalyse tussen de 100 en 136.

Onze resultaten lieten goede 'intra- en interobserver' betrouwbaarheid zien voor bijna alle scoringsmethoden. 'Observer' betrouwbaarheid voor categoriale data werd geanalyseerd met behulp van lineair gewogen kappa statistiek. Continue data werd geanalyseerd met behulp van de Intraclass Correlatie Coëfficiënt (ICC). De gecombineerde BASRI scoringsmethoden (BASR-s) en BASRI-t) en in het bijzonder de SASSS lieten een zeer goede betrouwbaarheid zien (ICC 0.85-0.98). Zelfs met een scoringsinterval van twee jaar was de 'intra-observer' betrouwbaarheid goed (ICC 0.85-0.96). De betrouwbaarheid van de relatief nieuwe scoringsmethode voor de heupen (BASRI-h) bleek ook goed te zijn (kappa 0.59-0.60).

Kappa geeft aan in welke mate twee 'observers' verschillen tussen röntgenfoto's kunnen waarnemen. Kappa is dus laag bij een homogene groep data waarbij iedere röntgenfoto dus min of meet dezelfde score heeft gekregen. Dit zou een verklaring kunnen zijn voor de relatief lage 'intra- en interobserver' betrouwbaarheid gevonden voor de SI scoringsmethoden (kappa 0.36-0.70). In deze situatie zijn statistische methoden die de geobserveerde overeenstemming relateren aan de verwachte overeenstemming (zoals kappa en ICC) van beperkte waarde gezien de hoge mate van verwachte overeenstemming. Dit wordt nog eens bevestigd door de relatief lage mediaan scores voor de SASSS scoringsmethoden (mediaan SASSS-totaal 17.5-18.0, mediaan 'gemodificeerde' SASSS 16.3-16.8, range 0-72). Verder geldt dat vanwege de lage prevalentie van radiologisch schade bij de SASSS de ICC toeneemt met als gevolg overschatting van de ICC.

Vanwege deze overwegingen ten aanzien van kappa en ICC hebben we ook concordanties berekend. De volledige concordantie tussen de twee 'observers' was over het algemeen laag (21-76%). Visuele presentatie van de continue data (SASSS scoringsmethoden) door middel van een Bland and Altman plot geeft ook meer inzicht in de data zeker omdat deze plots de verdeling van de data met 'outliners' over de gehele range van geobserveerde data laat zien. Deze plots lieten een maximaal verschil van 26 punten (range 0-72) tussen de twee 'observers' zien.

BASRI-s en BASRI-t waren de enige scoringsmethoden die verandering lieten zien bij een klein aantal patiënten na twee jaar follow-up (respectievelijk 7.5% en 7.4%) gebaseerd op een binomiaal afkappunt. Deze verandering kon niet worden aangetoond met de andere globale of gedetailleerde scoringsmethoden. De 'observers' kwamen in maximaal 52% overeen dat er geen verandering in BASRI-s en BASRI-t score plaats vond. Toch moeten we

helaas concluderen dat er zelden sprake was van relevante verandering omdat de 'observers' maar in 7.5% van de patiënten overeenstemde dat een significante verandering van minimaal 1 graad had plaats gevonden. Dit laatste kan misleidend zijn aangezien de definitie van 1 graad verandering een arbitrair afkappunt is. Het berekende kleinste aantoonbare verschil (SDD) gedefinieerd als afkappunt voor de SASSS scoringsmethoden is relatief groter. Het afkappunt gebruikt voor de SASSS methoden is dus strenger gedefinieerd dan het afkappunt (1graad verandering) gebruikt voor de BASRI methoden. Dit laatste zou weer een reden kunnen zijn waarom we geen verandering hebben gevonden in SASSS scores. Een mogelijke andere reden is dat 'observer variatie' of 'observer error' niet te onderscheiden is van radiologische progressie in onze studie. Het gebruik van binomiale afkappunten kan betekenen dat men informatie verliest en dat het daardoor niet meer mogelijk is verschillen aan te tonen. Een andere overweging is dat we een niet geselecteerde populatie Bechterew patiënten gebruikt hebben zonder een gegarandeerde mate van ziekte activiteit. In een groep Bechterew patiënten met een hoge ziekte activiteit zouden de resultaten wel eens anders kunnen zijn.
In onze studie is de betrouwbaarheid van de radiologische scoringsmethoden bij de ziekte van Bechterew goed. Helaas waren de scoringsmethoden onder de gegeven scorings condities (gepaard scoren met onbekende chronologische volgorde, resultaten gebaseerd op gemiddelde scores van twee 'observers', binomiale afkappunten en een niet geselecteerde patiënten populatie) niet in staat om na twee jaar follow-up bij een groot aantal patiënten radiologische progressie of verandering aan te tonen.

Resumé en perspectief

De ASAS werkgroep is opgericht vanwege gebrek aan uniformiteit op het gebied van ziekte-activiteit en ziekte-uitkomst in wetenschappelijke studies met betrekking tot de ziekte van Bechterew. Deze werkgroep heeft verschillende domeinen gedefinieerd met kernpunten voor drie verschillende onderzoeksettings bij de ziekte van Bechterew ('DC-ART', 'SM-ARD'/'physical therapy 'en'clinical record keeping'). In navolging op het werk van de ASAS werkgroep hebben wereldwijd diverse onderzoeksgroepen de laatste vijf jaar veel werk verricht door zich voornamelijk te richten op aspecten van ziekte-activiteit en ziekte-uitkomst bij de ziekte van Bechterew. De resultaten in dit proefschrift zijn voor het grootste deel afkomstig van de gegevens van een groot cohort Bechterew patiënten die longitudinaal gevolgd zijn (OASIS). Deze resultaten zijn gericht op aspecten van zowel ziekte-activiteit als ziekte-uitkomst. Hoofdstuk 6 en 7 gaan vooral over uitkomstmaten bij de ziekte van Bechterew waarbij hoofdstuk 6 de ontwikkeling en validatie van een ziekte specifieke kwaliteit van leven vragenlijst (ASQoL) beschrijft. Hoofdstuk 7 beschrijft de vergelijking van alle beschikbare radiologische scoringsmethoden bij de ziekte van Bechterew. De resultaten van deze studie laten zien dat de methoden betrouwbaar zijn maar dat geen van de methoden in staat is bij een groot aantal patiënten radiologische verandering aan te tonen

over een periode van twee jaar. In de nabije toekomst wordt waarschijnlijk duidelijk welke rol 'Magnetic Resonance Imaging' (MRI) gaat krijgen bij het aantonen van structurele verandering bij de ziekte van Bechterew. Het is met behulp van de MRI mogelijk om aspecten van schade en van ziekte-activiteit aan te tonen op het niveau van zowel de SI gewrichten als van de wervelkolom. Het mogelijke effect van bijvoorbeeld een bekende chronologische volgorde van röntgenfoto's op het aantonen van radiologische verandering bij conventionele radiologie wordt nog onderzocht.

Hoofdstuk 2 tot en met 5 beschrijven aspecten van ziekte-activiteit en de resultaten gepresenteerd in deze hoofdstukken bevestigen dat het meten van ziekte-activiteit bij de ziekte van Bechterew erg moeilijk blijft. Hoewel wereldwijd veel werk is gedaan, is er nog steeds geen uniforme maat die alle aspecten van ziekte-activiteit reflecteert. Acute fase reacties gemeten met CRP en BSE zijn slechts bij een minderheid van de Bechterew patiënten verhoogd. Verder blijkt dat Bechterew patiënten en hun behandeld artsen een heel ander zicht op ziekte-activiteit hebben. Tot nu toe wordt om deze reden ziekte-activiteit vanuit het patiënten perspectief, zoals BASDAI, in combinatie met ziekte-activiteit vanuit het perspectief van de arts en/of verhoogde CRP gebruikt als 'gouden standaard' voor het includeren van patiënten in klinische trials en voor de besluitvorming rondom het starten van anti TNF α therapie in de klinische praktijk. Anti TNF α en andere nieuwe biologische geneesmiddelen zijn recent beschikbaar gekomen bij de behandeling van de ziekte van Bechterew en de effecten lijken veelbelovend. In de toekomst zullen studies die de effecten van deze potente geneesmiddelen evalueren gebruikt kunnen worden voor validatie en vergelijking van de door de ASAS geselecteerde meetinstrumenten. Hopelijk leiden de resultaten van deze studies uiteindelijk naar de ontwikkeling van een uniform en valide meetinstrument dat alle aspecten van ziekte-activiteit bij de ziekte van Bechterew reflecteert.